holmusk

Word Embeddings Medical Domain

Presenter

Pranshu Chourasia

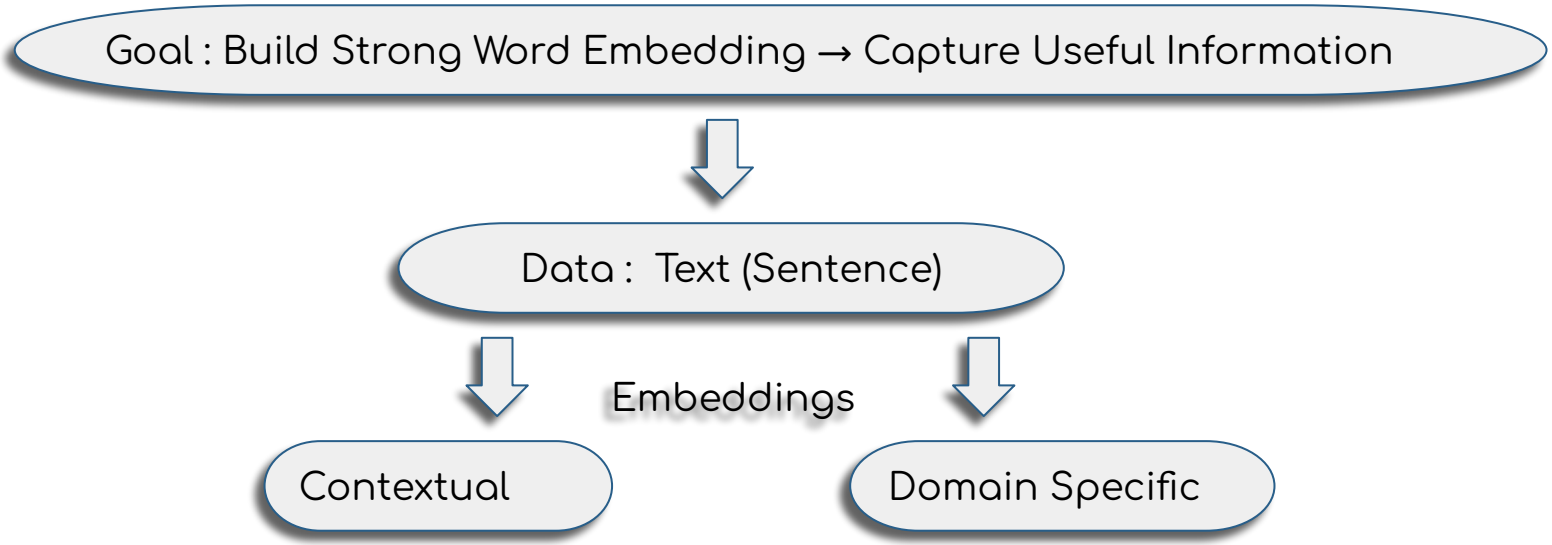
Table of contents

01

Use Case & Dataset Info.

- Understanding the use case.
- Exploratory Data Analysis on Datasets.

High Level Understanding



Dataset Information

MedicalConcepts.csv

- Two Columns :
  - Term1 & Term2
  - Closeness of pair
- Counts :
  - Total : 566 rows
  - Distinct : 558 rows
- Occurence of Term words in notes (Text) - 2255 times

ClinNotes.csv

- Two Columns :
  - Category & Notes
  - Notes (Text) → 3 Category
- Total Counts : 818 Records
- Occurence of words in Notes
  - Total : 373370
  - Distinct : 32553

Three Broader Vision

Classification :
Model Training on notes (Text) for classification of category

Information Extraction :
Fetching Entity information from notes (Text)

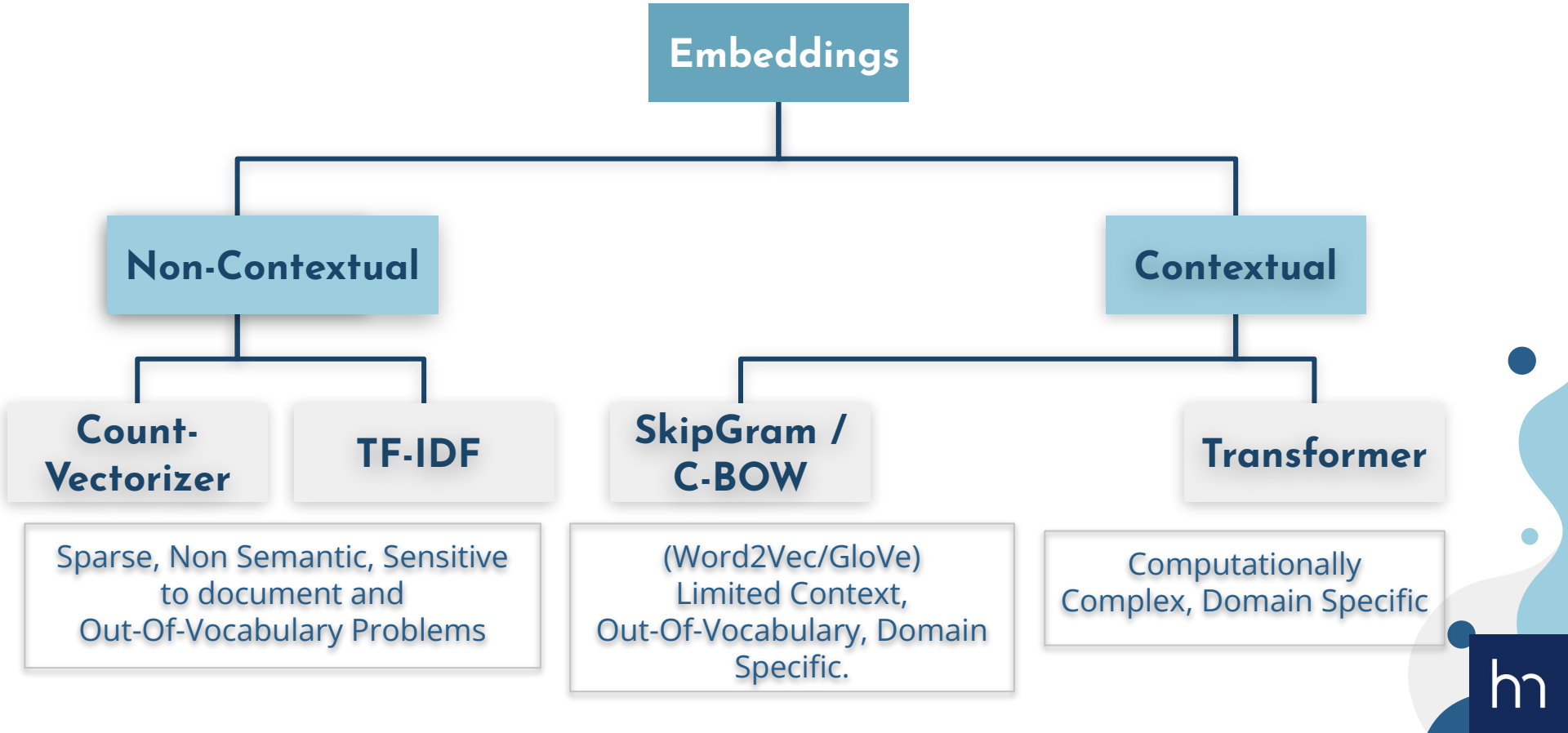
Knowledge :
High Vector Similarity of the Medical Concept Pairs

02

Embeddings & Model Research

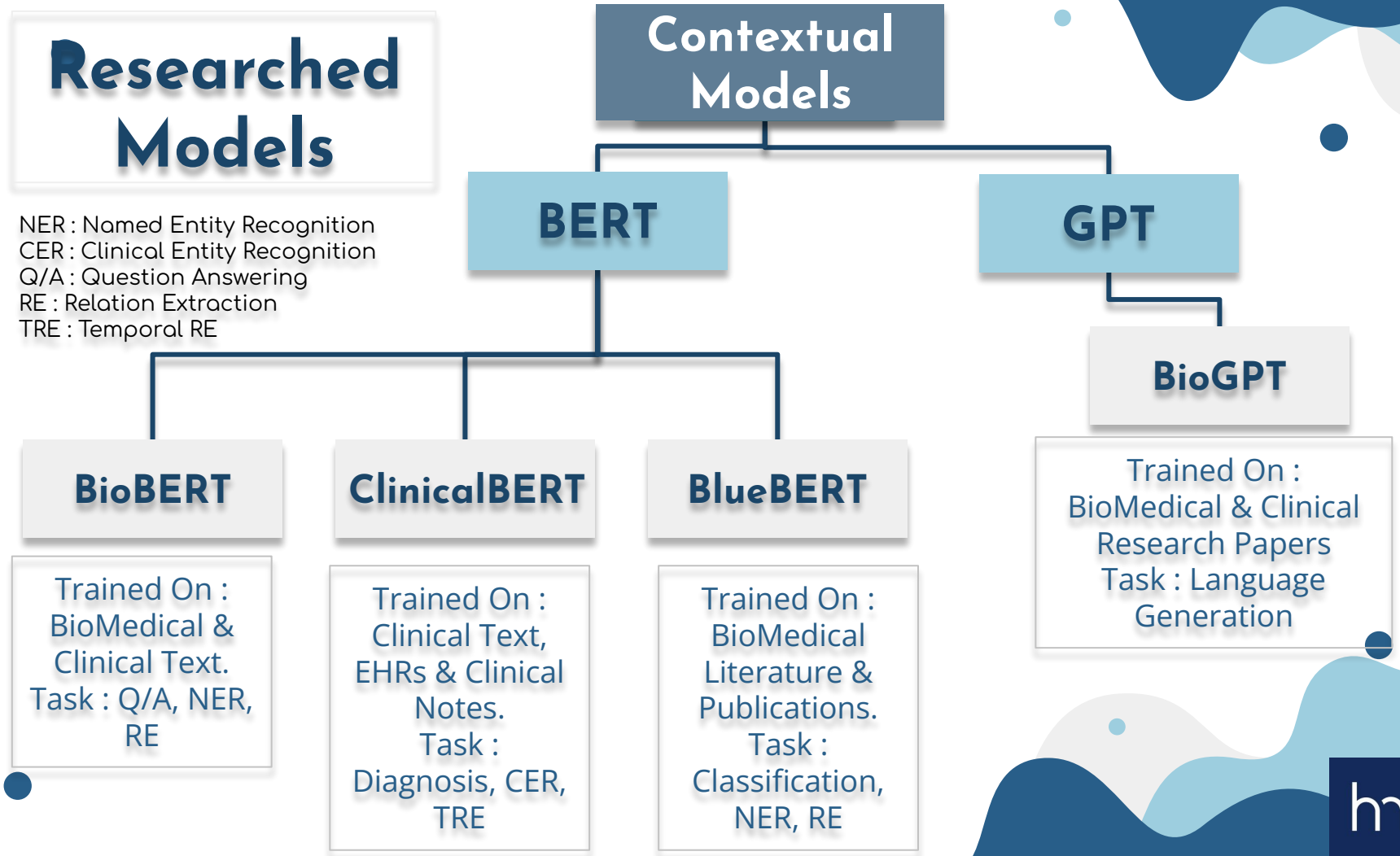
- Exploring different Word Embeddings.
- Researching Appropriate NLP Model.

Word Embedding



Researched Models

- NER : Named Entity Recognition
- CER : Clinical Entity Recognition
- Q/A : Question Answering
- RE : Relation Extraction
- TRE : Temporal RE



03

Model Tune & Explanations

- Generating Performance Metrics.
- Explaining Predictions (LIME)

Text (Notes) Preprocessing

Goal : Sentence Classification - Classify Notes to appropriate labels

Using SpaCy Module

- Lemmatization
- Stop Word removal
- Punctuation removal
- Empty String removal & Lower Case
- Custom processing

Additionally Perform Named Entity Recognition using SpaCy

Pre-Processed Text Examples

	category	category_label	notes	notes_preprocess
0	Cardiovascular / Pulmonary	0	2-D M-MODE: , ,1. Left atrial enlargement with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left ventricular ejection fraction of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology of aortic valve, mitral valve, tricuspid valve, and pulmonary valve.,7. PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral and tricuspid regurgitation.,2. Trace aortic and pulmonary regurgitation.	2 d m mode 1 left atrial enlargement left atrial diameter 4.7 cm 2 normal size right leave ventricle 3 normal lv systolic function leave ventricular ejection fraction 51%.4 normal lv diastolic function 5 pericardial effusion 6 normal morphology aortic valve mitral valve tricuspid valve pulmonary valve 7 pa systolic pressure 36 mmhg doppler 1 mild mitral tricuspid regurgitation 2 trace aortic pulmonary regurgitation

(a) Description for CardioVascular/Pulmonary - Label 0

308	Gastroenterology	1	PROCEDURE IN DETAIL: , Following a barium enema prep and lidocaine ointment to the rectal vault, perirectal inspection and rectal exam were normal. The Olympus video colonoscope then introduced into the rectum and passed by directed vision to the distal descending colon. Withdrawal notes an otherwise normal descending, rectosigmoid and rectum. Retroflexion noted no abnormality of the internal ring. No hemorrhoids were noted. Withdrawal from the patient terminated the procedure.	procedure detail follow barium enema prep lidocaine ointment rectal vault perirectal inspection rectal exam normal olympus video colonoscope introduce rectum pass direct vision distal descend colon withdrawal note normal descending rectosigmoid rectum retroflexion note abnormality internal ring hemorrhoid note withdrawal patient terminate procedure

(b) Description for Gastroenterology - Label 1

	Category	Notes	Category_label	Ner_Raw
114	Neurology	CC: ,Gait difficulty.,HX: ,This 59 y/o RHF was admitted with complaint of gait difficulty. The evening prior to admission she noted sudden onset of LUE and LLE weakness. She felt she favored her right leg, but did not fall when walking. She denied any associated dysarthria, facial weakness, chest pain, SOB, visual change, HA, nausea or vomiting.,PMH:, tonsillectomy, adenoidectomy, skull fx 1954, HTN, HA.,MEDS: ,none on day of exam.,SHX: ,editorial assistant at newspaper, 40pk-yr Tobacco, no	2	[['Gait', 'PERSON'], ['59', 'CARDINAL'], ['RHF', 'ORG'], ['LUE', 'ORG'], ['LLE', 'ORG'], ['SOB', 'ORG'], ['adenoidectomy', 'GPE'], ['skull fx 1954', 'ORG'], ['day', 'DATE'], ['40pk', 'ORDINAL'], ['etoh/drug', 'ORG'], ['ADMIT', 'ORG'], ['P95 R20', 'PERSON'], ['T36.6', 'ORG'], ['naming', 'ORG'], ['4/4', 'CARDINAL'], ['2/2', 'CARDINAL'], ['Fundi', 'ORG'], ['VFFTC', 'ORG'], ['Tongue ML', 'ORG'], ['hallucis longus', 'ORG'], ['Biceps/Triceps/Wrist', 'ORG'], ['LUE', 'ORG'], ['LLE', 'ORG'], ['Coord'...

Custom Processing

Replacing ‘y/o’ with ‘year old’

Notes_preprocess	Ner_processed
cc gait difficulty hx 59 year old rhf admit complaint gait difficulty evening prior admission note sudden onset lue lle weakness feel favor right leg fall walk deny associated dysarthria facial weakness chest pain sob visual change ha nausea vomiting pmh tonsillectomy adenoidectomy skull fx 1954 htn ha.meds day exam shx editorial assistant newspaper 40pk yr tobacco etoh drug fhx noncontributory admit exam p95 r20 t36.6 bp169/104ms a&o person place time speech fluent dysarthria naming compreh...	[['Gait', 'PERSON'], ['59 year old', 'DATE'], ['RHF', 'ORG'], ['LUE', 'ORG'], ['LLE', 'ORG'], ['SOB', 'ORG'], ['adenoidectomy', 'GPE'], ['skull fx 1954', 'ORG'], ['day', 'DATE'], ['40pk', 'ORDINAL'], ['etoh/drug', 'ORG'], ['ADMIT', 'ORG'], ['P95 R20', 'PERSON'], ['T36.6', 'ORG'], ['naming', 'ORG'], ['4/4', 'CARDINAL'], ['2/2', 'CARDINAL'], ['Fundi', 'ORG'], ['VFFTC', 'ORG'], ['Tongue ML', 'ORG'], ['hallucis longus', 'ORG'], ['Biceps/Triceps/Wrist', 'ORG'], ['LUE', 'ORG'], ['LLE', 'ORG'], ['...

Model Training

Split ClinNotes.csv Data (818 records) into 80% train and 20% test

Analysis on 4 models:

1. BioBERT
2. ClinicalBERT
3. BlueBERT
4. BioGPT

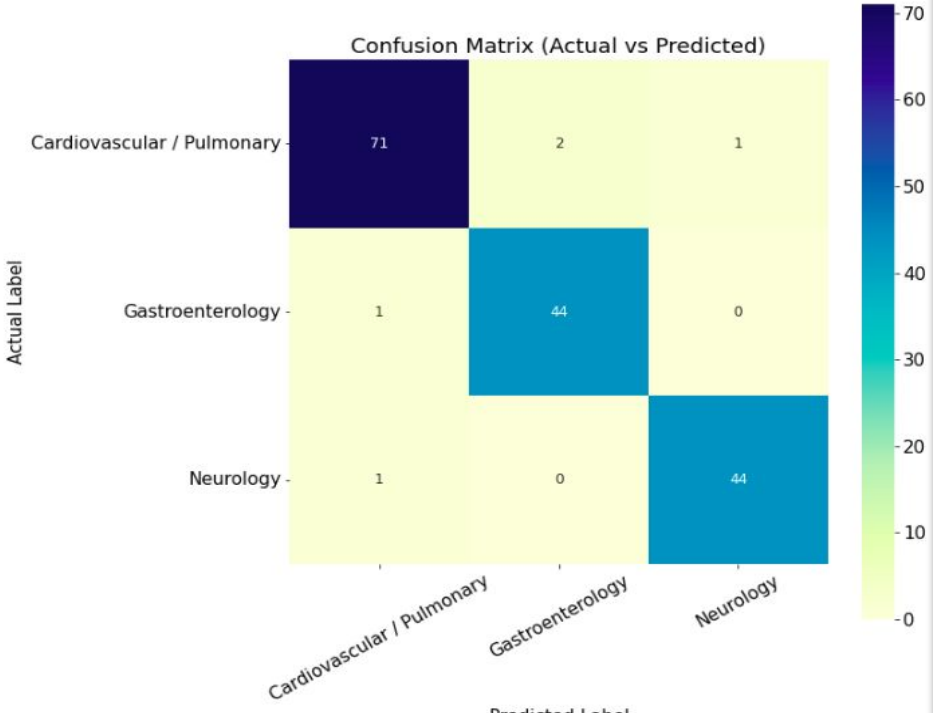
Download Pre-Trained Model - Huggingface

- Fine Tune each model on Raw text and later on Preprocessed Text
- Generated Performance Metrics
- Confusion Matrix & Classification Report
- Generate LIME explanations

ClinicalBERT Raw VS Preprocessed

ClinicalBERT_Raw

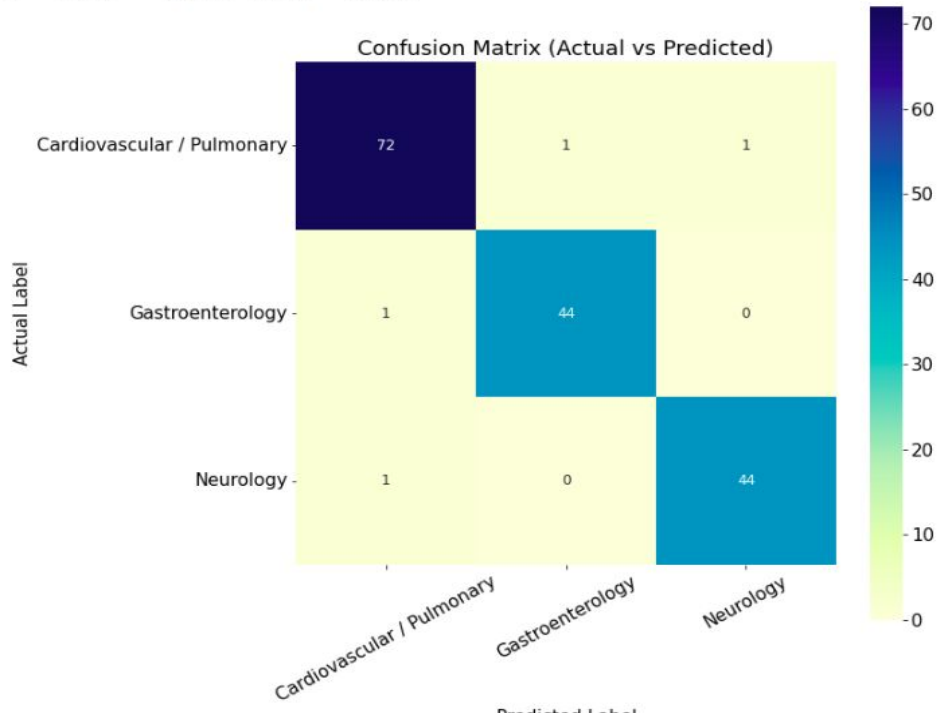
	Accuracy	Precision	Recall	F1-Score
0	96.951	96.961	96.951	96.951



	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.972603	0.959459	0.965986	74.000000
Gastroenterology	0.956522	0.977778	0.967033	45.000000
Neurology	0.977778	0.977778	0.977778	45.000000
accuracy	0.969512	0.969512	0.969512	0.969512
macro avg	0.968967	0.971672	0.970266	164.000000
weighted avg	0.969610	0.969512	0.969509	164.000000

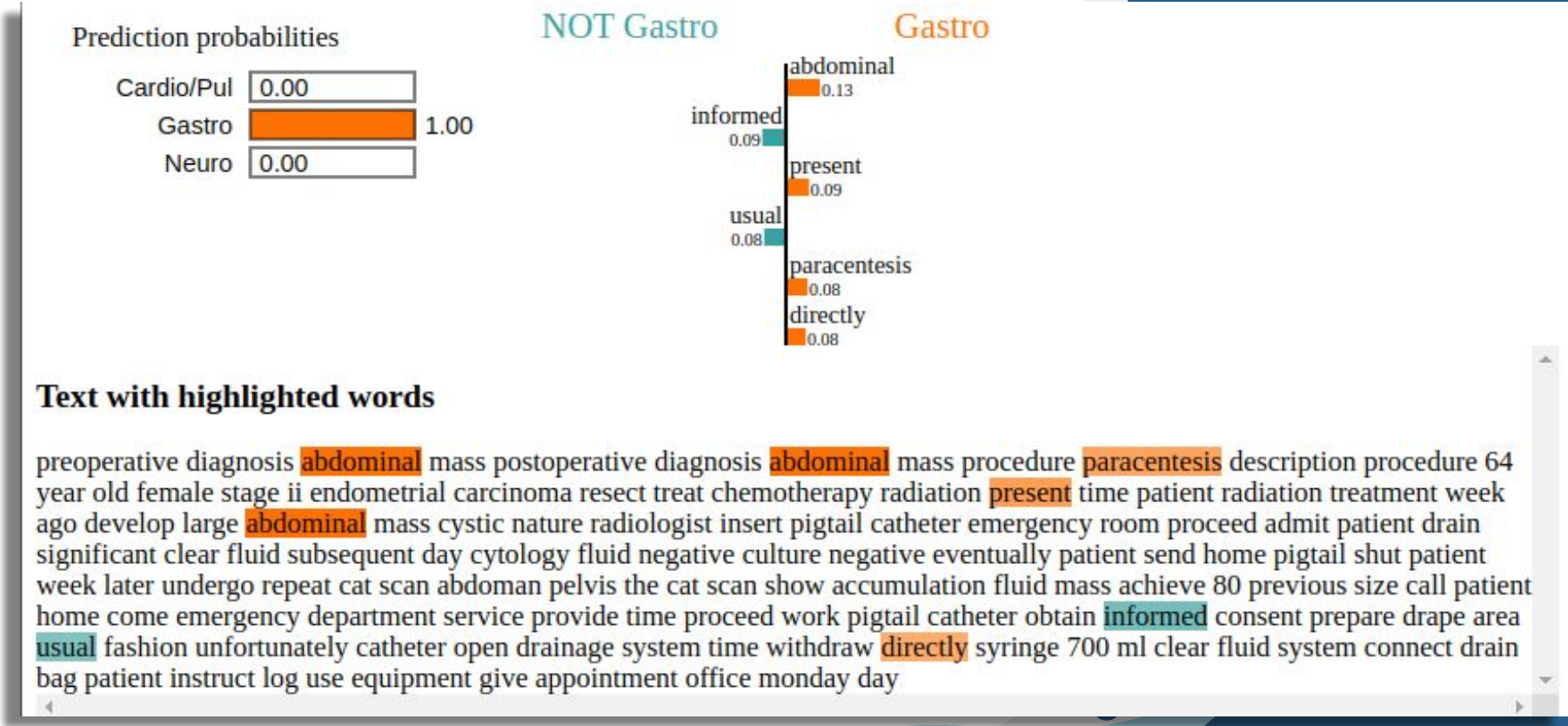
ClinicalBERT_Preprocess

	Accuracy	Precision	Recall	F1-Score
0	97.561	97.561	97.561	97.561



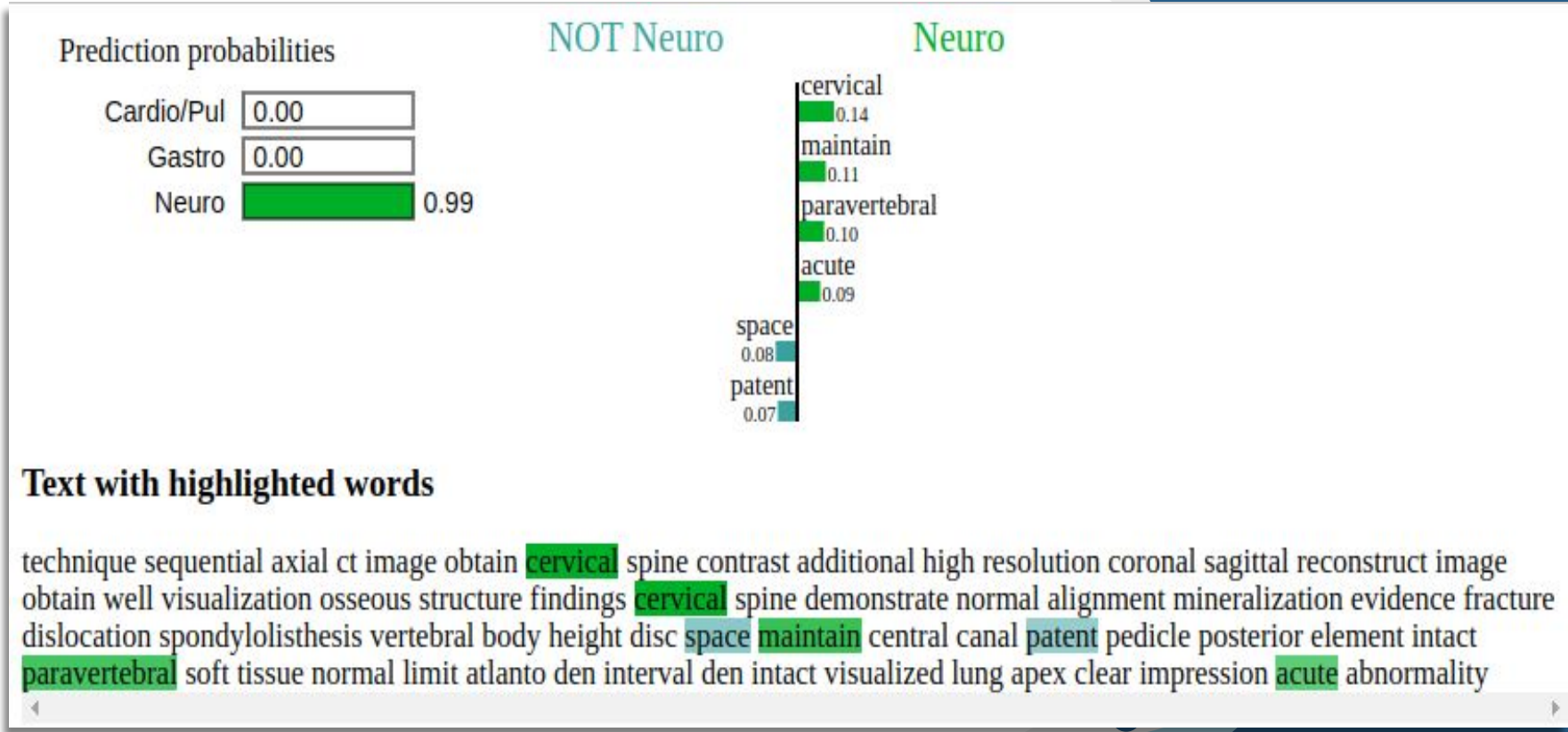
	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.972973	0.972973	0.972973	74.00000
Gastroenterology	0.977778	0.977778	0.977778	45.00000
Neurology	0.977778	0.977778	0.977778	45.00000
accuracy	0.975610	0.975610	0.975610	0.97561
macro avg	0.976176	0.976176	0.976176	164.00000
weighted avg	0.975610	0.975610	0.975610	164.00000

Lime Explanation - 1



(a) BioBERT Preprocessed (Gastroenterology - Label 1)

Lime Explanation - 2



(a) ClinicalBERT Preprocessed (Neurology - Label 2)

OverAll Model Results %

Models	Accuracy	Precision	Recall	F1-Score
BioBERT Raw	97.561	97.561	97.561	97.561
BioBERT Preprocess	97.561	97.561	97.561	97.561
ClinicalBERT Raw	96.951	96.961	96.951	96.951
ClinicalBERT Preprocess	97.561	97.561	97.561	97.561
BlueBERT Raw	96.951	96.961	96.951	96.951
BlueBERT Preprocess	96.951	96.964	96.951	96.951
BioGPT Raw	87.195	87.692	87.195	87.234
BioGPT Preprocess	93.902	94.059	93.902	93.859s

Mean of Cosine Similarity (%) of Embeddings of Term Pairs

	Base	Fine-Tune Raw	Fine-Tune Processed
BioBERT	82.06	58.06	53.00
ClinicalBERT	82.63	69.29	59.82
BlueBERT	73.66	53.85	50.12
BioGPT	92.66	99.57	99.52

04

Future Steps

- Understanding Existing Problem.
- Probable Solutions to Overcome.

Challenges

- BioGPT performance is not the best for Sentence Classification.
- BioGPT on term matching produces very high Similarity Scores

- BioBERT, ClinicalBERT and BlueBERT's performance is almost similar for Sentence Classification Task
- Fine Tuned model performance not upto the mark for term similarity

- For Knowledge Extraction very basic SpaCy Model was used.

Probable Solutions

Knowledge Transfer from Term Pairs

- Use customized loss function which incorporates cosine distance, while fine tuning.
- Train simpler model on term pairs (word2vec) and produces those as custom embeddings to BERT.

BioBERT Embedding trained over classifier

- Use BioBERT model to generate sentence embeddings, later use those to train a classifier.
- Overcome lack of Data Problem

Advanced Models For Knowledge Extraction

- Advanced Pre trained BERT models will perform better for NER, CER, etc knowledge extraction task than simple spaCy Model.

Thanks

Do you have any questions?

+91 9990266468